Making LLMs Fight Until They Agree: A User-Driven Protocol for Cross-Model Validation and Hallucination Reduction

Abstract

Large language models frequently produce confident but incomplete, contradictory, or hallucinated outputs. While academic work has explored multi-agent debate and self-critique, everyday users lack a simple, repeatable method to verify answers without defaulting to any single model.

This paper presents Adversarial Convergence, a practical protocol in which the user prompts multiple frontier models with the same question, then forces the models to directly critique and confront one another's responses in successive rounds until substantial agreement is reached or irreducible disagreements are clearly isolated.

Developed and deployed by the author during pro se federal litigation in 2025 against OpenAI and related parties, the method was used to validate arguments and exhibits under real legal pressure. It has since been refined through daily use across multiple domains. The protocol requires no advanced technical skills and can be applied immediately by any user with access to several models.

The goal is to empower individuals to move beyond passive acceptance of single-model outputs and toward independent verification.

Introduction

In 2025, as a pro se plaintiff in federal court facing OpenAI, Microsoft, and Sam Altman, I needed reliable answers on complex legal and factual matters. Relying on any single model proved too risky. Models regularly hallucinated facts, contradicted themselves, or softened positions when challenged, a known issue called sycophancy. Alignment techniques such

as RLHF often optimize for user appeasement and plausible-sounding answers rather than strict accuracy.

To counter this, I developed a method of forcing multiple models to confront one another directly. I collected initial responses, then required each model to critique the others, iterating until convergence or clear mapping of disagreements. The approach proved effective enough that converged outputs were submitted as court exhibits. I have continued using and refining the protocol daily since.

This paper documents the method as a practical tool for any user. It is not a theoretical academic paper but a field-tested framework built out of necessity.

Origin and Timeline

The protocol originated in early 2025 during preparation for Bullens v. OpenAI L.P. et al. in the U.S. District Court for the Southern District of Indiana. Early experiments involved two or three models. By mid-2025 the workflow had matured into a repeatable process. Converged outputs supported court filings. The method has remained in active daily use through 2026.

The Adversarial Convergence Protocol

Step 1. Define the question clearly and neutrally. Avoid leading language.

Step 2. Collect initial responses from at least three frontier models using the identical query.

Step 3. Launch confrontation rounds. Paste all responses into each model and instruct it to critique the others for factual errors, logical gaps, unsupported claims, and inconsistencies.

Step 4. Iterate by feeding the new critiques back to all models. Continue until termination criteria are met.

Protocol Termination Metrics (The Verification Threshold)

The workflow uses explicit stopping conditions to maintain integrity. It terminates under one of two conditions:

1. Semantic Convergence (80 percent matrix): Core assertions, timelines, and conclusions reach an estimated 80 percent parity across all models for two consecutive rounds.

2. Irreducible Dialectical Split: Disagreements are no longer due to hallucinations or errors but reflect fundamental differences in training data or perspectives. At this point the user must resolve the split manually.

Example prompts for confrontation rounds include:

“Here are responses from other models on the same question. Identify factual errors, logical gaps, unsupported claims, or inconsistencies. Explain your reasoning and provide your revised answer.”

“Compare all current answers. List points of agreement and disagreement. Propose a unified response that resolves contradictions.”

Why the Method Works

Forcing models to defend their outputs against peer critique counters sycophancy and single-model optimization for plausibility. Different architectures and training data create natural checks. Iterative confrontation surfaces weaknesses that remain hidden in isolated queries. The result is a form of practical cross-validation that improves reliability for the user.

Real-World Results

During the 2025 litigation the method helped identify and correct potential errors before filing. In daily use since then it has supported technical analysis, document review, news evaluation, and personal decisions. Public demonstrations on X, including direct challenges that brought Gemini into conversation with Grok, have shown consistent convergence on stronger answers.

Limitations

The protocol requires access to multiple models and more time than a single query. It does not eliminate all bias if models share similar training weaknesses. Highly subjective topics benefit less. Rate limits and costs can constrain heavy use. Human oversight remains essential.

Comparison to Existing Work

Academic literature includes multi-agent debate and critique frameworks. This protocol distinguishes itself through real-world deployment in federal litigation in 2025 and ongoing refinement by an independent user focused on practical accessibility.

How Anyone Can Start Today

Select any three available models. Begin with a low-stakes question you can independently verify. Follow the steps above. Keep a simple log of the process. With practice the workflow becomes efficient and intuitive.

Conclusion

Users should not accept any single AI response as final truth. The Adversarial Convergence protocol offers a straightforward way to verify outputs by making models confront one another until agreement emerges or disagreements are clearly isolated. Developed under legal pressure and refined through daily application, it provides a practical tool for sovereign reasoning.

Anyone can adopt this method today. The result is greater confidence in AI-assisted work and a meaningful step toward independent verification.

Appendix

Starter confrontation prompt:

“Question: [insert question]

I will provide responses from other models. Critique them thoroughly for errors or weaknesses, then give your best revised answer.”

Users are encouraged to experiment, document results, and share refinements while preserving the core confrontation-and-convergence structure.